IMPROVING CORRUPTION ROBUSTNESS WITH ADVERSARIAL FEATURE ALIGNMENT TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite their success, vision transformers still remain vulnerable to image corruptions, such as noise or blur. Indeed, we find that the vulnerability mainly stems from the unstable self-attention mechanism, which is inherently built upon patch-based inputs and often becomes overly sensitive to the corruptions across patches. For example, when we only occlude a small number of patches with random noise (e.g., 10%), these patch-based corruptions would lead to severe accuracy drops and greatly mislead the intermediate features as well as the corresponding attentions over them. To alleviate this issue, we seek to explicitly reduce the sensitivity of attention layers to patch-based corruptions and improve the overall robustness of transformers. In this paper, we propose the **Adversarial Feature Alignment Transformer (AFAT)** that aligns the features between clean examples and patch-based corruptions. To construct these corrupted examples, we build a patch corruption model to identify and occlude the patches that could severely distract the intermediate attention layers. We highlight that the corruption model is trained adversarially to the following feature alignment process, which is essentially different from existing methods. In experiments, AFAT greatly improves the stability of attention layers and consistently yields better robustness on various benchmarks, including CIFAR-10/100-C, ImageNet-A, ImageNet-C, and ImageNet-P.

1 INTRODUCTION

Despite the success of vision transformers (Dosovitskiy et al., 2021) in recent years, they still lack robustness against common image corruptions (Hendrycks & Dietterich, 2019; Mao et al., 2022) such as noise or blur. For example, even for the state-of-the-art robust architectures, e.g., RVT (Mao et al., 2022) and FAN (Zhou et al., 2022), the accuracy drops by more than 15% on corrupted examples, e.g., with Gaussian noise, as shown in Figure 1 (blue star on the right). We suspect that this vulnerability is inherent to the used self-attention mechanism, which relies on patch-based input and may easily become overly sensitive to corruptions or perturbations upon them. In this paper, we aim to develop a more reliable self-attention scheme and start our journey by first studying the stability of intermediate attention layers against *patch-based* corruptions.

For example, as illustrated in Figure 1, we construct patch-based corruptions by occluding a small number of patches with random noise, e.g., 10%. Considering RVT (Mao et al., 2022) as a strong baseline, these corrupted patches have significant impact on the attention maps across layers, as shown in Figure 1 (left). We suspect this to be the case due to the global interactions across tokens in the self-attention mechanism – even when occluding or corrupting only few patches. Quantitatively, this can be captured by computing the average cosine similarity between the attentions on clean and corrupted images across layers, denoted by Cos-Sim. Regarding the considered example in Figure 1, the Cos-Sim of only 0.43 for RVT indicates a significant shift in attention – a phenomenon that we can observe across the entire validation set of ImageNet (see Figure 4). In fact, these attention shifts also have direct and severe impact on accuracy: In Figure 1 (right), we randomly sample 100 occlusion masks for each image and show the distribution of accuracy (blue box). Unsurprisingly, the accuracy decreases significantly when facing the patch-based corruptions, compared to the full/original examples (blue star). These experiments highlight the need for an inherently more robust attention mechanism in order to improve the overall robustness of transformers.

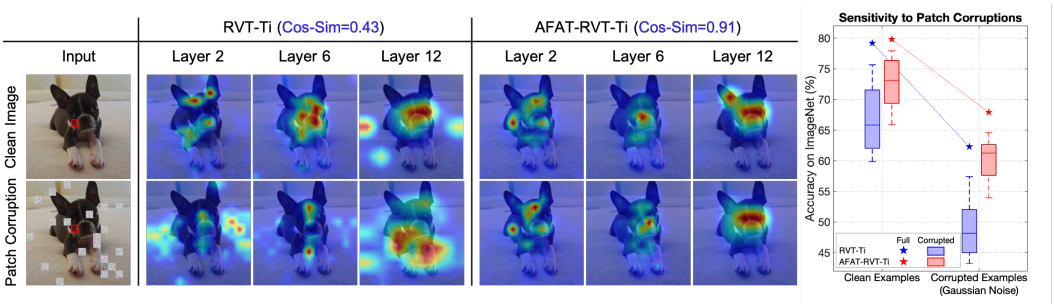

Figure 1: Sensitivity to patch-based corruptions in terms of attention stability (left) and accuracy (right). *Left*: We randomly occlude 10% patches with noise and show the attention maps of different layers in RVT-Ti (Mao et al., 2022) and our AFAT-RVT-Ti. Following (Fu et al., 2022), we choose the center patch (red square) as the query and average the attention scores across all the attention heads for visualization. Regarding this example, we also compute the average cosine similarity (Cos-Sim) between the clean and corrupted attentions across different layers. Clearly, our AFAT model yields more stable attention maps. *Right*: On ImageNet, we plot the distribution of accuracy on the occluded examples with different occlusion masks. Here, we randomly sample 100 different masks for each image. We show that RVT is very sensitive to the patch-based corruptions and has a much larger variance of accuracy than our AFAT model.

We address this problem by finding particularly vulnerable patches to be occluded/corrupted and then explicitly stabilizing the intermediate self-attention layers against them. As shown in Figure 1 (right), with a fixed occlusion ratio, the accuracy still varies a lot when occluding different patches (e.g., ranging from 60% to 75% in the blue box). Since we seek to stabilize the attention mechanism of transformers, occluding the most vulnerable (often very important) patches and explicitly reducing the impact of them should bring the largest robustness improvement. Inspired by this, we seek to identify these vulnerable patches to construct patch-based corruptions and then align the intermediate features to make the attention less sensitive to the corruptions in individual patches, making the overall model more robust. In practice, we are able to reduce the impact of patch-based corruptions significantly, improving the Cos-Sim from 0.43 (for RVT) to 0.91 in Figure 1 (left). This is also directly observed in the visual results where these corruptions have little impact on the intermediate attention maps of our AFAT model. The stable attention mechanism also greatly improves the robustness of transformers. As shown in Figure 1 (right), compared with RVT, we obtain significantly higher accuracy when facing examples with different occlusion masks (red box), alongside the improved overall accuracy and robustness on full images (red star).

Contributions: In this paper, we propose the **Adversarial Feature Alignment Transformer (AFAT)**, making three key contributions: *(1)* We propose an *adversarial feature alignment* scheme that adversarially selects patches to be occluded/corrupted and then explicitly aligns the intermediate features between clean and patch-based corrupted examples. *(2)* To construct these patch-based corruptions, we develop a patch corruption model that finds particularly vulnerable patches by maximizing the distance of intermediate features between clean and corrupted examples. In practice, the patch corruption model is trained adversarially to the classification model and greatly benefits the following feature alignment process. *(3)* In experiments, we demonstrate that the robustness improvement against patch-based corruptions (shown in Figure 1 (right)) can generalize well to various robustness benchmarks, including ImageNet-A/C/P (Zhao et al., 2018; Hendrycks & Dietterich, 2019). More critically, we can show, both qualitatively and quantitatively, that these improvements stem from the more stable attention mechanism across layers.

2 Related Work

Vision Transformers (ViTs) (Dosovitskiy et al., 2021; Vaswani et al., 2017; Heo et al., 2021; Wang et al., 2021b; Yuan et al., 2021) have achieved remarkable performance in various learning tasks. Besides improving clean accuracy, many works seek to study and improve the robustness of ViTs (Bhojanapalli et al., 2021; Bai et al., 2021; Shi et al., 2020; Paul & Chen, 2022; Benz et al., 2021b). In-

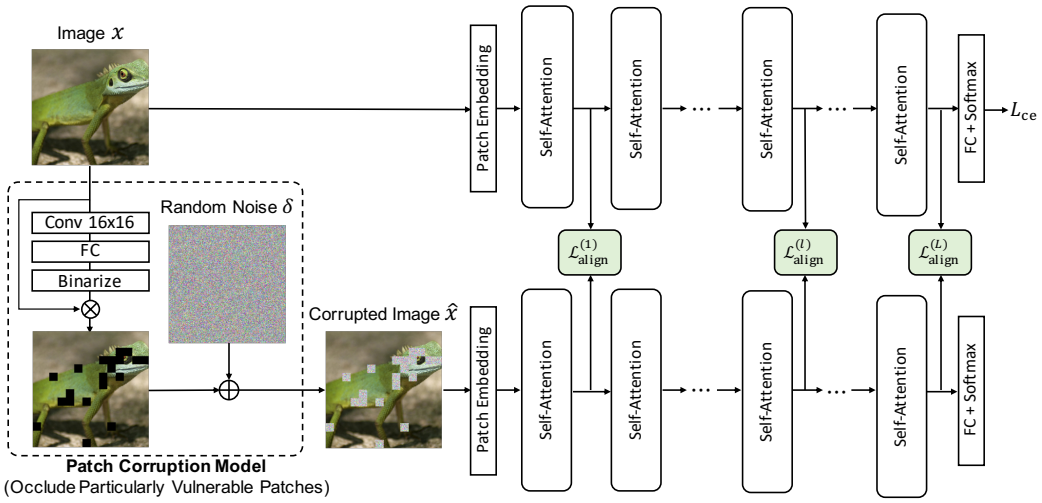

Figure 2: Overview of the Adversarial Feature Alignment Transformer (AFAT). We present a patch corruption model to produce patch-based corruptions and align the features of each self-attention block between the clean and corrupted examples (the alignment loss is highlighted by green box). Unlike existing methods, we select the patches to be occluded/corrupted in an adversarial way, i.e., corrupting the most vulnerable patches that would greatly distract intermediate attention layers.

terestingly, ViTs are often more robust than convolutional networks against corruptions (Qin et al., 2022; Tang et al., 2021; Tian et al., 2022; Gu et al., 2022) and adversarial attacks (Mahmood et al., 2021; Shao et al., 2021; Naseer et al., 2021; Lovisotto et al., 2022; Shi & Han, 2021; Liang et al., 2022; Benz et al., 2021a). To further improve robustness, RVT (Mao et al., 2022) develops a robust transformer by comparing different designs for each component, and presents a patch-wise augmentation. FAN (Zhou et al., 2022) combines token attention and channel attention (Ali et al., 2021) and yields new state-of-the-arts. However, even for these robust transformers, there is still a large gap between clean and robust accuracy. More critically, the impact of corruptions/perturbations on the key component of ViTs, i.e., self-attention, still remains poorly understood.

Besides the above, an intuitive way to improve robustness is to reduce the gap of intermediate features between clean and corrupted examples, e.g., using feature alignment (Chen et al., 2019; Wang et al., 2021a; Zhang et al., 2021; Burns & Steinhardt, 2021; Song et al., 2019; Yan et al., 2021). Typically, feature alignment aligns the features of examples from two different domains. However, when considering corruption robustness, the corruption type of test data is often unknown and the corrupted training examples are also unavailable. To tackle this, we focus on ViTs and develop a patch corruption model to produce corrupted examples for feature alignment. Unlike existing methods, we select the patches in an adversarial way, i.e., finding the most vulnerable patches, to introduce corruptions. Based on the generated examples with patch-based corruptions, we investigate the stability of self-attention modules and develop a robust feature alignment approach for ViTs.

3 Adversarial Feature Alignment Transformer (AFAT)

We suspect that the vulnerability of state-of-the-art transformers to corruptions stems from the inherently fragile self-attention mechanism. To address this issue, we propose the Adversarial Feature Alignment Transformer (AFAT) that explicitly enhances the stability of self-attention to improve the overall robustness. We achieve this goal by identifying the vulnerable patches to construct patch-based corruptions and then aligning their features with the clean ones at intermediate attention layers, as shown in Figure 2. As for the first step of AFAT, in Section 3.2, we propose a patch corruption model that finds particularly vulnerable patches to construct patch-based corruptions. Then, in Section 3.1, we develop an adversarial feature alignment scheme that jointly trains the corruption model and the classification model using an adversarial objective.

Algorithm 1 Training method for the **Adversarial Feature Alignment Transformers (AFATs)**. We train the classification model $\mathcal{F}$ and the corruption model $\mathcal{C}$ in an end-to-end manner. In each iteration, we descend the gradient for $\mathcal{F}$ and ascend the gradient for $\mathcal{C}$, respectively.

Require: Training data $\mathcal{D}$, model parameters $\theta_{\mathcal{C}}$ and $\theta_{\mathcal{F}}$, occlusion ratio ρ, step size η, hyper-parameter λ.

for each training iteration **do**
  Sample a data batch $\{x_i\}_{i=1}^N$ from $\mathcal{D}$
  // `Construct the patch-based corrupted example` $\hat{x}$
  Sample the random noise δ from a uniform distribution
  Construct $\hat{x}$ using the patch corruption model $\mathcal{C}$: $\hat{x} = \mathcal{C}(x;\rho)\cdot x + (1-\mathcal{C}(x;\rho))\cdot\delta$
  // `Update the classification model` $\mathcal{F}$
  Update $\theta_{\mathcal{F}}$ by descending the gradient: $\theta_{\mathcal{F}} = \theta_{\mathcal{F}} - \eta\frac{1}{N}\sum_{i=1}^N \nabla_{\theta_{\mathcal{F}}}\left[\mathcal{L}_{\text{ce}}(x_i) + \lambda\mathcal{L}_{\text{align}}(x_i,\hat{x}_i)\right]$
  // `Update the patch corruption model` $\mathcal{C}$
  Update $\theta_{\mathcal{C}}$ by ascending the gradient: $\theta_{\mathcal{C}} = \theta_{\mathcal{C}} + \eta\frac{1}{N}\sum_{i=1}^N \nabla_{\theta_{\mathcal{C}}}\lambda\mathcal{L}_{\text{align}}(x_i,\hat{x}_i)$
end for

3.1 Finding Vulnerable Patches to Construct Patch-based Corruptions

In the first step of AFAT, we build a patch corruption model $\mathcal{C}$ to construct the examples with patch-based corruptions that would severely distract the intermediate attention layers. As shown in Figure 2 (bottom left), given an example x and an occlusion ratio ρ, the corruption model first predicts a binary mask $M(x) = \mathcal{C}(x;\rho)$ to determine which patches should be occluded/corrupted. Then, we occlude them with a random noise δ sampled from the uniform distribution. Formally, the patch-based corrupted example can be formulated by

$$\hat{x} = M(x)\cdot x + (1-M(x))\cdot\delta, \quad \text{where } M(x) = \mathcal{C}(x;\rho). \tag{1}$$

We highlight that, compared with simply dropping patches, occluding patches with noise is more challenging for the model and particularly effective in practice (as detailed in Table 5 in appendix).

As shown in Figure 1 (right, blue box), randomly occluding/corrupting patches often incurs significantly different impacts on accuracy between the best and worst case. This leads us to consider what patches should be occluded to perform effective feature alignment. Since we seek to improve the stability against them, we propose to construct the worst case by finding those vulnerable patches that, once occluded, can greatly distract the intermediate attention layers. To this end, we train the patch corruption model by maximizing the distance of intermediate features between clean and corrupted examples. Let $\mathcal{F}_l(x)$ be the features obtained at the l-th layer for x. Given a model with L layers, the training objective of the patch corruption model becomes

$$\max_{\mathcal{C}} \ \mathbb{E}_{x\sim\mathcal{D}}\ \mathcal{L}_{\text{align}}(x,\hat{x}), \quad \text{where } \mathcal{L}_{\text{align}}(x,\hat{x}) = \frac{1}{L}\sum_{l=1}^{L}\|\mathcal{F}_l(x) - \mathcal{F}_l(\hat{x})\|^2. \tag{2}$$

Here, $\mathcal{D}$ denotes the distribution of data and $\mathcal{L}_{\text{align}}(x,\hat{x})$ denotes the feature alignment loss that measures the average feature distance over all the attention layers. Compared with directly maximizing the cross-entropy loss, maximizing $\mathcal{L}_{\text{align}}$ explicitly distracts the attention layers and in practice performing alignment against it brings larger robustness improvement (see Table 6 in appendix).

Architecture of the patch corruption model. As shown in Figure 2 (bottom left), the corruption model is a lightweight network that contains a convolution followed by a fully connected layer and a binarization layer. The binarization layer is essentially a (hard) threshold function that selects the top ρ of the patches to be occluded and keeps the rest unchanged. Following (Hubara et al., 2016), we use the Straight Through Estimator (STE) to make the binarization operation differentiable.

3.2 Adversarial Feature Alignment

Regarding the second step of AFAT, as shown in Figure 2, we seek to stabilize the self-attention layers by aligning the intermediate features between clean and patch-based corrupted examples. To make sure that we can always construct the most challenging corrupted examples w.r.t. the latest classification model, we simultaneously train the corruption model $\mathcal{C}$ and the classification model $\mathcal{F}$ using an adversarial objective. Specifically, we minimize both the cross-entropy loss $\mathcal{L}_{\text{ce}}(x)$ and the alignment loss $\mathcal{L}_{\text{align}}(x,\hat{x})$ to train $\mathcal{F}$, while maximizing the alignment loss for $\mathcal{C}$. Since $\mathcal{L}_{\text{ce}}(x)$ only relies on the clean example, our training objective can be equivalently formulated as

$$\min_{\mathcal{F}}\max_{\mathcal{C}}\ \mathbb{E}_{x\sim\mathcal{D}}[\mathcal{L}_{\text{ce}}(x) + \lambda\mathcal{L}_{\text{align}}(x,\hat{x})], \tag{3}$$

Table 1: Comparisons with the state-of-the-art on CIFAR-10 and CIFAR-100. We evaluate clean accuracy on the original test set and robust accuracy on the corresponding corrupted datasets, i.e., CIFAR-10-C and CIFAR-100-C. We show that our AFAT significantly improves the robustness on both datasets. † denotes models with the same training recipe as used for our AFAT.

	Model	#Params (M)	CIFAR-10 (%)	CIFAR-10-C (%)	CIFAR-100 (%)	CIFAR-100-C (%)
CNN	ResNet50	23.5	94.77	84.81	76.43	66.75
	ResNet50†	23.5	96.01	87.53	81.16	68.08
	RLAT	11.7	94.75	89.60	-	-
	PRIME	11.7	93.06	89.05	77.60	68.28
	NoisyMix	6.1	96.73	92.78	81.16	72.06
	AdA	23.5	94.93	92.17	-	-
	CARD-Decks	44.6	96.80	92.75	80.60	71.30
ViT	Swin-T	27.6	95.84	90.25	81.83	70.87
	DeiT-S	21.7	95.30	89.01	79.84	68.79
	ConViT-S	27.4	96.90	91.73	82.43	71.39
	RVT-S	23.0	97.21	92.35 (+0.00)	84.13	73.43 (+0.00)
	+ AFAT (Ours)	23.0	97.73	94.14 (+1.79)	84.81	74.94 (+1.51)
	FAN-S-Hybrid	25.7	97.69	93.14 (+0.00)	84.92	74.19 (+0.00)
	+ AFAT (Ours)	25.7	**98.06**	**94.59 (+1.45)**	**85.30**	**75.72 (+1.53)**

where λ is a trade-off parameter determining the importance of $\mathcal{L}_{\text{align}}$. The impact of λ on clean accuracy and robustness can be found in Figure 6. Regarding Eqn. (3), we highlight that constructing examples to perform feature alignment in an adversarial way greatly improves the robustness and is essentially different from existing feature alignment approaches.

To solve the minimax problem (3), we update the models $\mathcal{F}$ and $\mathcal{C}$ by descending and ascending the gradients, respectively. As shown in Algorithm 1, the training procedure has two key steps. First, we produce the corrupted examples $\hat{x}$ using $\mathcal{C}$ and descend the gradient w.r.t. Eqn. (3) to update the parameters $\theta_{\mathcal{F}}$ of the classification model $\mathcal{F}$. Second, we ascend the gradients to update the parameters $\theta_{\mathcal{C}}$ of the patch corruption model $\mathcal{C}$, enforcing it to produce the worst-case corrupted examples. When ascending the gradient, we can directly change the sign of gradients to update the parameters, making it possible for end-to-end training.

4 Experiments

We conduct extensive experiments to evaluate our AFAT approach based on two state-of-the-art robust architectures, including RVT (Mao et al., 2022) and FAN (Zhou et al., 2022). In Section 4.1, we first justify our method on the CIFAR datasets and show that AFAT achieves new state-of-the-arts on two corruption test sets, namely CIFAR-10-C and CIFAR-100-C. Then, in Section 4.2, we perform comparisons on ImageNet and demonstrate that AFAT greatly improves the robustness on various robustness benchmarks, including ImageNet-A, ImageNet-C, and ImageNet-P.

4.1 Comparisons on CIFAR-10 and CIFAR-100

In this experiment, we train the models from scratch on CIFAR-10/100 and compare both accuracy and corruption robustness. Following (Calian et al., 2022), we use DeepAugment (Hendrycks et al., 2020) and train the models for 200 epochs. We adopt the batch size of 128 and use cosine decay to adjust the learning rate. For fair comparisons, we consider RVT-S (Mao et al., 2022) and FAN-S-Hybrid (Zhou et al., 2022) as the baselines that contain approximately the same number of parameters with popular CNNs and transformers. In all the experiments, by default, we set $\lambda{=}5{\times}10^{-3}$ and $\rho{=}10\%$ for our AFAT models. Please see more ablations on these hyper-parameters in Section 5.

In Table 1, we compare our AFATs with both state-of-the-art CNNs (Kireev et al., 2022; Modas et al., 2021; Erichson et al., 2022; Calian et al., 2022; Diffenderfer et al., 2021; Guo et al., 2022) and popular transformer models (Liu et al., 2021; Touvron et al., 2021; d'Ascoli et al., 2021; Mao et al., 2022; Zhou et al., 2022). To make fair comparisons with CNNs, we also apply the training recipe of transformers to train a ResNet50 model, denoted by ResNet50† in Table 1. Specifically, we do not exploit our patch corruption model or feature alignment, but directly apply the same augmentation for training. Compared with CNNs, transformers tend to obtain higher accuracy but do

Table 2: Comparisons of robustness on ImageNet. We report the mean corruption error (mCE) on ImageNet-C and mean flip rate (mFR) on ImageNet-P. The lower mCE or mFR is, the more robust the model is. Across different model sizes, our AFAT models consistently improve the robustness.

	Model	#FLOPs (G)	#Params (M)	ImageNet	Robustness Benchmarks			
					IN-A	IN-C ↓	IN-C w/o Noise ↓	IN-P ↓
CNN	ResNet50	4.1	25.6	76.1	0.0	76.7	76.0	58.0
	Inception v3	5.7	27.2	77.4	10.0	80.6	82.0	61.3
	ANT	4.1	25.6	76.1	1.1	63.0	64.3	53.2
	EWS	4.1	25.6	77.3	5.9	58.7	60.2	30.9
	DeepAugment	4.1	25.6	75.8	3.9	60.6	52.2	32.1
ViT-Tiny	Deit-Ti	1.3	5.7	72.2	7.3	71.1	72.9	56.7
	ConViT-Ti	1.4	5.7	73.3	8.9	68.4	70.4	53.7
	PiT-Ti	0.7	4.9	72.9	6.2	69.1	70.8	53.2
	PVT-Tiny	1.9	13.2	75.0	7.9	69.1	70.0	60.1
	RVT-Ti	1.3	10.9	79.2	14.6 (+0.0)	57.0 (-0.0)	58.9 (-0.0)	39.1 (-0.0)
	+ AFAT (Ours)	1.3	10.9	79.5	16.5 (+1.9)	**55.7 (-1.3)**	**57.5 (-1.4)**	38.0 (-1.1)
	FAN-T-Hybrid	3.5	7.5	80.1	21.9 (+0.0)	58.3 (-0.0)	59.8 (-0.0)	38.3 (-0.0)
	+ AFAT (Ours)	3.5	7.5	**80.3**	**23.6 (+1.7)**	57.2 (-1.1)	58.4 (-1.4)	**37.3 (-1.0)**
ViT-Small	Deit-S	4.6	22.1	79.9	6.3	54.6	56.6	36.9
	ConViT-S	5.4	27.8	81.5	18.9	49.8	52.1	35.8
	Swin-T	4.5	28.3	81.2	21.6	62.0	64.2	38.3
	PiT-S	2.9	23.5	80.9	21.7	52.5	54.7	36.7
	PVT-Small	3.8	24.5	79.9	18.0	66.9	70.0	45.1
	T2T-ViT_t-14	6.1	21.5	81.7	23.9	53.2	54.4	36.2
	RVT-S	4.7	23.3	81.9	25.7 (+0.0)	49.4 (-0.0)	51.6 (-0.0)	35.2 (-0.0)
	+ AFAT (Ours)	4.7	23.3	82.2	27.9 (+2.2)	48.4 (-1.0)	50.4 (-1.2)	34.3 (-0.9)
	FAN-S-Hybrid	6.7	25.7	83.5	33.9 (+0.0)	48.5 (-0.0)	50.7 (-0.0)	34.5 (-0.0)
	+ AFAT (Ours)	6.7	25.7	**83.6**	**36.8 (+2.9)**	**47.5 (-1.0)**	**49.4 (-1.3)**	**33.5 (-1.0)**
ViT-Base	Deit-B	17.6	86.6	82.0	27.4	48.5	50.9	32.1
	ConViT-B	17.7	86.5	82.4	29.0	46.9	49.3	32.2
	Swin-B	15.4	87.8	83.4	35.8	54.4	57.0	32.7
	PiT-B	12.5	73.8	82.4	33.9	48.2	51.0	32.3
	PVT-Large	9.8	61.4	81.7	26.6	59.8	63.0	39.3
	T2T-ViT_t-24	15.0	64.1	82.6	28.9	48.0	49.3	31.8
	RVT-B	17.7	91.8	82.6	28.5 (+0.0)	46.8 (-0.0)	49.8 (-0.0)	31.9 (-0.0)
	+ AFAT (Ours)	17.7	91.8	82.8	32.1 (+3.6)	45.7 (-1.1)	48.5 (-1.3)	31.0 (-0.8)
	FAN-B-Hybrid	11.3	50.5	83.9	39.6 (+0.0)	46.1 (-0.0)	48.1 (-0.0)	31.3 (-0.0)
	+ AFAT (Ours)	11.3	50.5	**84.2**	**41.1 (+1.5)**	**44.5 (-1.6)**	**46.8 (-1.3)**	**30.0 (-1.2)**

not necessarily exhibit better robustness, such as Swin (Liu et al., 2021), DeiT (Touvron et al., 2021), and ConViT (d'Ascoli et al., 2021). As for the carefully designed robust architecture RVT (Mao et al., 2022) and FAN (Zhou et al., 2022), they both greatly improve the robustness and outperform existing methods in most cases. Compared with the RVT and FAN baselines, our AFAT models further improve the corruption robustness by a large margin, i.e., with the improvement larger than 1.4% on both CIFAR-10-C and CIFAR-100-C. More critically, our AFAT-FAN-S-Hybrid modes achieve new state-of-the-art results on both benchmarks.

4.2 Comparisons on ImageNet

On ImageNet, we evaluate our method based on both RVT (Mao et al., 2022) and FAN (Zhou et al., 2022). We consider three variants of each architecture according to the model size, including Tiny, Small, and Base. Again, we closely follow the settings of RVT and FAN for training. Specifically, we train the models using the same augmentation schemes and adopt the batch size of 2048. We set the learning rate to 2×10^{-3} and train all the models for 300 epochs. To evaluate the robustness, we consider several robustness benchmarks, including ImageNet-A (IN-A) (Zhao et al., 2018), ImageNet-C (Hendrycks & Dietterich, 2019), and ImageNet-P (IN-P) (Hendrycks & Dietterich, 2019). Since we also introduce noise to construct the patch-based corruptions, we also report the results on IN-C without the corruption types related to noise (i.e., excluding Gaussian Noise, Shot Noise, and Impulse Noise from the 15 corruption types). Following (Hendrycks & Dietterich, 2019), we report the mean corruption error (mCE) on IN-C (also IN-C w/o Noise) and mean flip rate (mFR) on IN-P. For both metrics, *lower is better*. Regarding the two considered baseline architectures, for fair comparisons, we use the official evaluation code of RVT (Mao et al., 2022) to evaluate the pre-trained models as well as our AFAT models on these robustness benchmarks.

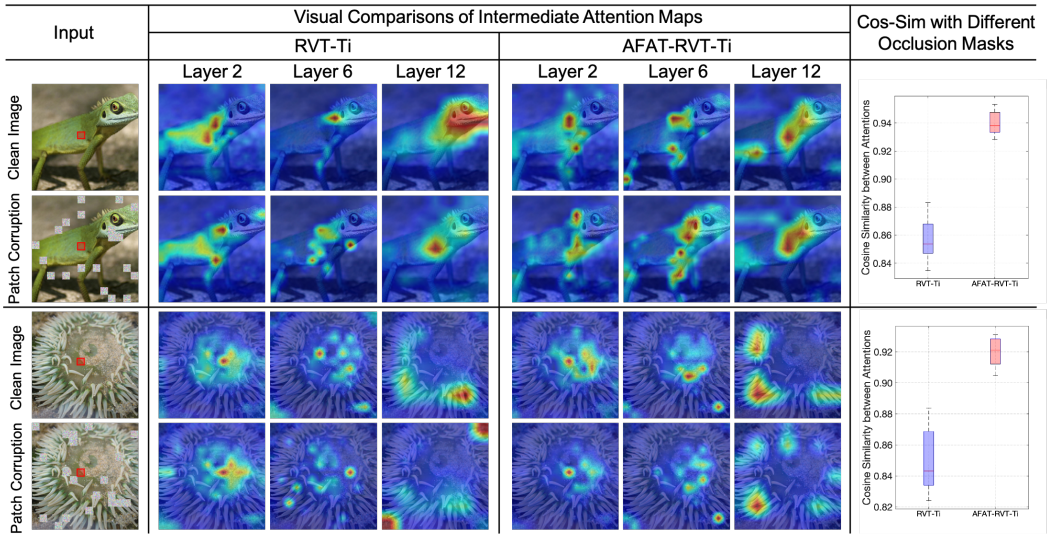

Figure 3: Comparisons of attention stability between RVT-Ti and our AFAT-Ti. We adopt the same method as that in Figure 1 to obtain the attention maps for visualization. In the last column, we also investigate the impact of different occlusion masks (1000 random masks) on each example and quantitatively evaluate the stability using cosine similarity (Cos-Sim). Clearly, our AFAT yields much more stable attention maps both qualitatively and quantitatively.

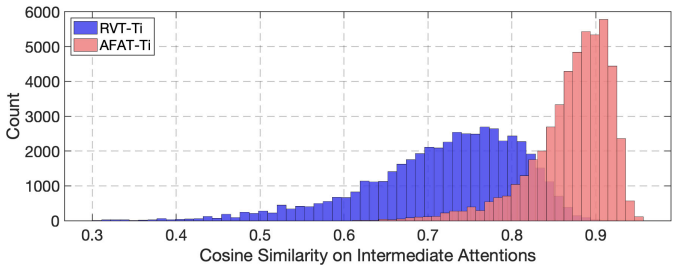

Figure 4: Histogram of cosine similarity on intermediate attention maps on ImageNet. For each image, we construct the corrupted example using a random occlusion mask and compute the average cosine similarity across layers. Clearly, AFAT yields much more stable attentions than RVT.

As shown in Table 2, compared with the strong baselines RVT and FAN, AFAT models consistently improve the robust accuracy on IN-A by $>1.5\%$ across different model sizes while keeping comparable clean accuracy. Moreover, we reduce the corruption error by $>1.0\%$ on IN-C and by $>1.2\%$ on IN-C without noise corruption types, which indicates that our method not only improves the robustness on noise corruptions but also generalizes well to the other corruption types. Moreover, the robustness on individual corruption type can be found in Figure 7 of appendix. When evaluating the stability against perturbations on IN-P, our AFAT models also show clear superiority over the RVT and FAN baselines. Overall, these experiments indicate that performing adversarial feature alignment consistently improves robustness across different architectures and corruption types.

5 Analysis and Discussions

In the following, we present further analysis on attention stability and visualize the generated patch-based corruptions in Section 5.1. In Section 5.2, we report the robustness against both patch-based corruptions and adversarial attacks. Moreover, we also discuss several hyper-parameters in AFAT, including the occlusion ratio ρ and the weight of alignment loss λ.

5.1 Visualization Results and More Analysis

Stability of intermediate attention maps. In this experiment, we directly visualize how much the intermediate attentions would be changed when facing patch-based corruptions. We take RVT-Ti as the baseline and compare the attention maps between RVT-Ti and our AFAT-RVT-Ti. Following (Fu et al., 2022), we average the attention maps across all the attention heads in each layer and visualize

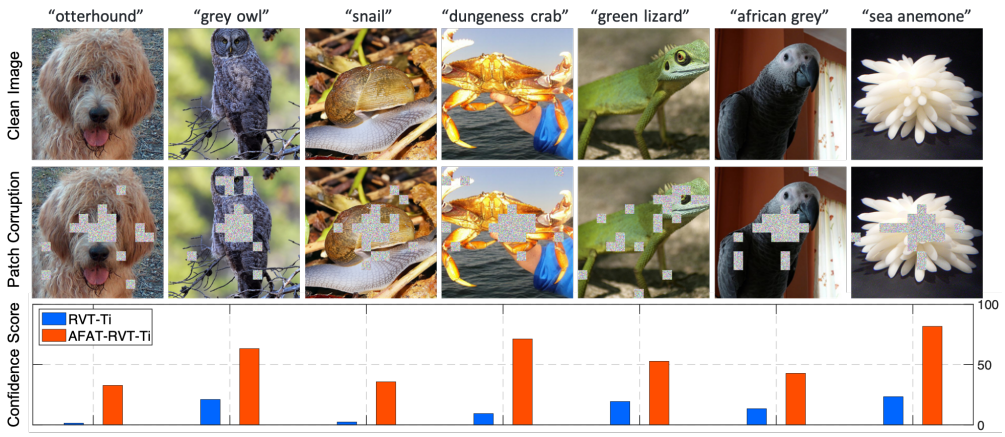

Figure 5: Visualization of the patch-based corrupted examples produced by the proposed patch corruption model (the second row) and comparisons of confidence score (bottom row). The corruption model often identifies those patches that would significantly distract intermediate self-attention layers if they are occluded. Moreover, our AFAT model tends to yield higher confidence score even when facing the generated patch-based corruptions.

Table 3: Comparisons of robustness against patch-based corruptions with the occlusion ratio $\rho = 10\%$, patch-based perturbations (e.g., Patch-Fool (Fu et al., 2022)), and PGD attack on ImageNet. Our AFAT models consistently outperforms the baseline models against them.

Method		RVT-B	FAN-B-Hybrid
Patch-based Corruption (Random Selected Patch)	Vanilla	73.0 (+0.0)	73.7 (+0.0)
	AFAT	**75.0 (+2.0)**	**75.2 (+1.5)**
Patch-based Corruption (Generated by $\mathcal{C}$)	Vanilla	61.4 (+0.0)	62.9 (+0.0)
	AFAT	**64.1 (+2.7)**	**65.2 (+2.3)**
Patch-Fool	Vanilla	69.3 (+0.0)	71.2 (+0.0)
	AFAT	**70.9 (+1.6)**	**72.5 (+1.3)**
PGD-5	Vanilla	29.9 (+0.0)	30.5 (+0.0)
	AFAT	**30.8 (+0.9)**	**31.7 (+1.2)**

the attention map for a query token, e.g., the center token highlighted by the red box. In Figure 3, we show that RVT often incurs significant changes in the attention maps. By contrast, our AFAT effectively preserves most of the regions with relatively high attention scores across layers. We also quantitatively evaluate the attention stability by computing the *cosine similarity* between the attention maps extracted from the clean and patch-based corrupted examples. Here, we compute the cosine similarity for each head in all the layers and then report the average score over them. Figure 4 plots histograms of attention similarity scores across the whole validation set of ImageNet. Clearly, AFAT increases the similarity both on average and in the worst-case across the whole dataset. In addition, Figure 3 (last column) also studies impact of different occlusion masks and shows the distribution of this score for two example images, each with 1000 random occlusion masks.

Patch-based corruptions generated by $\mathcal{C}$. We also visualize the patch-based corruptions generated by our patch corruption model $\mathcal{C}$ in Figure 5. By maximizing the feature alignment loss, the patch corruption model is able to detect those patches that have major contributions in the attention mechanism but on the other hand would make the attention very unstable if they are corrupted/occluded. In practice, the patch corruption model tends to occlude the patches that are mainly located in the key part of the object, e.g., the eyes of the dog in the first example. In the bottom of this figure, we also compare the confidence score of the ground-truth class predicted by both RVT and our AFAT models. Clearly, the generated patch-based corruptions often greatly reduce the confidence score of RVT. By contrast, as we explicitly perform feature alignment against these patch corruptions, our AFAT model still yields promising confidence score and thus comes with better robustness.

5.2 More Results and Ablations

Robustness against patch-based corruptions and adversarial attacks. In this experiment, we evaluate the robustness against patch-based corruptions (with the corruptions on both randomly selected patches and adversarial selected patches) and adversarial attacks. Regarding adversarial robustness, we report the accuracy on Patch-Fool (Fu et al., 2022) and the standard PGD attack (Madry

Table 4: Comparisons of different strategies for feature alignment on ImageNet and ImageNet-C (IN-C). We take RVT-B and FAN-B-Hybrid as the baselines in this experiment. We show that the adversarial feature alignment yields significantly better robustness the random strategy while keeping comparable accuracy with the baseline without the alignment loss.

Patch Selection Strategy	RVT-B		FAN-B-Hybrid	
	Imagenet	IN-C ↓	Imagenet	IN-C ↓
Baseline (w/o $\mathcal{L}_{\text{align}}$)	82.6	46.8 (-0.0)	83.9	46.1 (-0.0)
AFAT with Random Patch Selection	82.7	46.5 (-0.3)	84.1	45.6 (-0.5)
AFAT with Adversarial Patch Selection	**82.8**	**45.7 (-1.1)**	**84.2**	**44.5 (-1.6)**

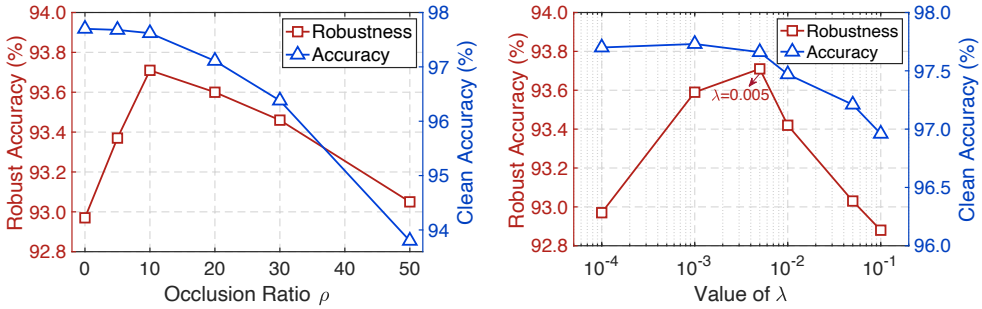

Figure 6: Accuracy and corruption robustness of AFAT-RVT-S on CIFAR-10 plotted against the occlusion ratio ρ (left) and the weight of the alignment loss (right). *Left:* $\rho = 10\%$ obtains the best robustness along with comparable accuracy. *Right:* A too small or too large value of λ hampers the robustness of AFAT. We found that $\lambda = 0.005$ performs best in most cases.

et al., 2018). Here, we consider the vanilla Patch-Fool (single patch, i.e. 1P) without constraints and follow the hyper-parameters of the original paper. As for PGD attack, we follow the settings of RVT (Mao et al., 2022) to construct the adversarial examples with the number of steps $t = 5$ and step size $\alpha = 0.5$. From Table 3, our AFAT models consistently outperform the baseline models on both patch-based corruptions and adversarial attacks.

Patch selection strategy. As mentioned in Section 3.1, we seek to find the vulnerable patches to be occluded/corrupted in an adversarial way. To justify this, we compare our method with the random patch selection strategy. As shown in Table 4, performing feature alignment with the adversarial patch selection strategy greatly outperforms the random strategy on ImageNet-C. This experiment indicates that adversarially selecting patches to introduce corruptions is particularly effective.

Occlusion ratio ρ and the weight of alignment loss λ. In this experiment, we train the AFAT-RVT-S model on CIFAR-10 datasets to investigate the impact of the occlusion ratio ρ and the weight of alignment loss λ. In Figure 6 (left), we change the value ρ by within the range between 0% and 50%. In practice, we obtain the best robustness with the ratio of 10% and thus set ρ=10% in our experiments. In Figure 6 (right), we discuss the weight of our adversarial alignment loss λ. Given a set of candidate values as shown in this figure, we observe that λ=0.005 yields the best robustness along with competitive clean accuracy. We highlight that these hyper-parameters also generalize well on CIFAR-100 and ImageNet. We adopt these settings in all the experiments.

6 Conclusion

In this paper, we study the robustness of transformer models by investigating the stability of self-attention against patch-based corruptions. For most self-attention modules, the features and the corresponding attentions over them are very sensitive to these corruptions, which, however, contributes to the lack of overall robustness. To alleviate this, we propose the Adversarial Feature Alignment Transformers (AFATs) that explicitly stabilize the intermediate features of each self-attention layer. Specifically, we develop an adversarial feature alignment approach that aligns the features between clean examples and the examples with patch-based corruptions. We highlight that introducing corruptions to those adversarially selected patches is particularly effective in feature alignment. In practice, AFAT greatly improves the stability of self-attention as well as the overall robustness.

REFERENCES

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. In *Proc. of the British Machine Vision Conference (BMVC)*, 2021a.

Philipp Benz, Chaoning Zhang, Soomin Ham, Adil Karjauv, and I Kweon. Robustness comparison of vision transformer and mlp-mixer to cnns. In *Proceedings of the CVPR 2021 Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV)*, pp. 21–24, 2021b.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 10231–10241, 2021.

Collin Burns and Jacob Steinhardt. Limitations of post-hoc feature alignment for robustness. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Dan A Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, Andras Gyorgy, Timothy Mann, and Sven Gowal. Defending against image corruptions through adversarial augmentations. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2022.

Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 627–636, 2019.

James Diffenderfer, Brian Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura. A winning hand: Compressing deep networks can improve out-of-distribution robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *Proc. of the International Conference on Machine Learning (ICML)*, pp. 2286–2296. PMLR, 2021.

N Benjamin Erichson, Soon Hoe Lim, Francisco Utrera, Winnie Xu, Ziang Cao, and Michael W Mahoney. Noisymix: Boosting robustness by combining data augmentations, stability training, and noise injections. *arXiv.org*, 2202.01263, 2022.

Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? In *Proc. of the International Conference on Learning Representations (ICLR)*, 2022.

Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *Proc. of the European Conference on Computer Vision (ECCV)*. Springer, 2022.

Yong Guo, David Stutz, and Bernt Schiele. Improving robustness by enhancing weak subnets. In *Proc. of the European Conference on Computer Vision (ECCV)*. Springer, 2022.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv.org*, abs/2006.16241, 2020.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 11936–11945, 2021.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 1012–1021. PMLR, 2022.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv.org*, 2202.07800, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, and Jan Hendrik Metzen. Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15234–15243, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Prime: A few primitives can boost robustness to common corruptions. *arXiv.org*, 2112.13547, 2021.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv.org*, 2106.04169, 2021.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proc. of the Conference on Artificial Intelligence (AAAI)*, volume 36, pp. 2071–2081, 2022.

Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv.org*, 2103.15670, 2021.

Yucheng Shi and Yahong Han. Decision-based black-box attack against vision transformers via patch-wise adversarial removal. *arXiv.org*, 2112.03492, 2021.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2020.

Chuanbiao Song, Kun He, Jiadong Lin, Liwei Wang, and John E Hopcroft. Robust local features for improving the generalization of adversarial training. *Proc. of the International Conference on Learning Representations (ICLR)*, 2019.

Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan L. Yuille, Philip H. S. Torr, and Dacheng Tao. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv.org*, abs/2109.05211, 2021.

Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, and Yugang Jiang. Deeper insights into vits robustness towards common corruptions. *arXiv.org*, 2204.12143, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. of the International Conference on Machine Learning (ICML)*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tao Wang, Ruixin Zhang, Xingyu Chen, Kai Zhao, Xiaolin Huang, Yuge Huang, Shaoxin Li, Jilin Li, and Feiyue Huang. Adaptive feature alignment for adversarial training. *arXiv.org*, 2105.15157, 2021a.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 568–578, 2021b.

Hanshu Yan, Jingfeng Zhang, Gang Niu, Jiashi Feng, Vincent Tan, and Masashi Sugiyama. Cifs: Improving adversarial robustness of cnns via channel-wise importance-based feature selection. In *Proc. of the International Conference on Machine Learning (ICML)*, pp. 11693–11703. PMLR, 2021.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 558–567, 2021.

Xiaoqin Zhang, Jinxin Wang, Tao Wang, Runhua Jiang, Jiawei Xu, and Li Zhao. Robust feature learning for adversarial defense via hierarchical feature alignment. *Information Sciences*, 560: 256–270, 2021.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *Proc. of the International Conference on Machine Learning (ICML)*, pp. 27378–27394. PMLR, 2022.

A Overview and Outline

In the main paper, we study the vulnerability of transformers by investigating the stability of self-attention against patch-based corruptions. Since the existing self-attention mechanism tend to be overly sensitive to patch-based corruptions, we propose the **Adversarial Feature Alignment Transformer (AFAT)**, which explicitly enhances the stability of self-attention against patch-based corruptions to improve the overall robustness. In the appendix, we provide additional ablations and complementary experimental results. We organize the appendix as follows:

- In Section B, we conduct additional ablation studies on the effect of occluding patches with noise and the effect of our feature alignment loss in generating patch-based corruptions. We show that, compared with dropping patches, occluding patches with noise greatly benefits the feature alignment process. Moreover, generating the corrupted examples using our feature alignment loss encourages the model to obtain larger robustness improvement than the cross-entropy loss.
- In Section C, we evaluate our AFAT models on each individual corruption type of ImageNet-C. We show that AFAT outperforms the baseline model in most corruption types, alongside the better overall robustness on the whole dataset.
- In Section D, we include more qualitative results to demonstrate the effectiveness of our AFAT in improving the stability of attention maps. Experiments show the superiority of our AFAT over the RVT baseline in improving the stability of attention maps.

B More Ablations and Discussions

In this section, we conduct further ablations of the proposed method based on RVT-B on ImageNet. Specifically, we first compare occluding patches with noise with simply dropping patches when generating the patch-based corruptions. Then, we also demonstrate the superiority of our feature alignment loss over the cross-entropy loss in identifying vulnerable patches.

Occluding patches with noise. As mentioned in the main paper, we construct the patch-based corruptions by occluding patches with noise. Actually, one can also directly drop these patches, namely PatchDrop. Here, we empirically compare these two methods based on RVT-B model on ImageNet(-C). As shown in Table 5, performing feature alignment against both PatchDrop and the occlusion with noise consistently improves the clean accuracy. Nevertheless, occluding patches with noise encourages the model to obtain better robustness than PatchDrop on ImageNet-C. The main reason is that, directly dropping patches often imposes relatively weak impact on the model and thus comes with limited performance improvement. By contrast, occluding patches with noise provides more severe impact on the intermediate features, making the feature alignment more challenging and also more effective.

Table 5: Comparisons between occluding patches with noise and dropping patches on ImageNet and ImageNet-C. We show that stabilizing the models against the occluded patches with noise yields significantly better robustness than stabilizing against PatchDrop.

Method	ImageNet ↑	ImageNet-C (mCE) ↓
Baseline (RVT-B)	82.6	46.8 (-0.0)
AFAT (PatchDrop)	**82.8**	46.3 (-0.5)
AFAT (Occlusion with Noise)	**82.8**	**45.7 (-1.1)**

Effect of feature alignment loss in generating patch-based corruptions. We further empirically compare the proposed feature alignment loss $\mathcal{L}_{\text{align}}$ with the cross-entropy loss in generating the patch-based corruptions. As shown in Eqn. (2), we seek to maximize $\mathcal{L}_{\text{align}}$ to identify those vulnerable patches that would greatly distract the intermediate attention layers. Indeed, we can also directly maximize the cross-entropy loss $\mathcal{L}_{\text{ce}}$ on the final prediction to perform patch selection. We compare these two approaches and show the results in Table 6. Clearly, maximizing $\mathcal{L}_{\text{align}}$ yields significantly better robustness than maximizing $\mathcal{L}_{\text{ce}}$. The main reason is that the cross-entropy loss only focuses on the final layer and may fail to mislead the intermediate attention layers. In contrast, maximizing $\mathcal{L}_{\text{align}}$ explicitly distracts all the attention layers in the model (also including the final layer) and would potentially encourage the whole model to be more robust when performing feature alignment against the generated patch corruptions.

Table 6: Comparisons between cross-entropy loss and feature alignment loss in generating patch-based corruptions based on RVT-B. Generating patch corruptions using the feature alignment loss encourages the model to obtain significantly better robustness than the cross-entropy loss.

Patch Selection Strategy	ImageNet ↑	ImageNet-C (mCE) ↓
Baseline (RVT-B)	82.6	46.8 (-0.0)
AFAT with Patch Selection ($\max \mathcal{L}_{\text{ce}}(\hat{x})$)	82.7	46.2 (-0.6)
AFAT with Patch Selection ($\max \mathcal{L}_{\text{align}}(x, \hat{x})$)	**82.8**	**45.7 (-1.1)**

C Robustness on Individual Corruption Type

In this experiment, we compare the corruption error on each individual corruption type of ImageNet-C between RVT-Ti and our AFAT-RVT-Ti. As shown in Figure 7, our AFAT model yields lower corruption error than the baseline model in most of the corruption types except the corruption type of motion blur. Although we introduce random noise when generating the patch-based corruptions, the major improvement of corruption robustness does not come from the noise related corruptions. Instead, the improved robustness can generalize well to other corruptions, e.g., yielding clearly lower error on the corruptions of snow, frost, and fog.

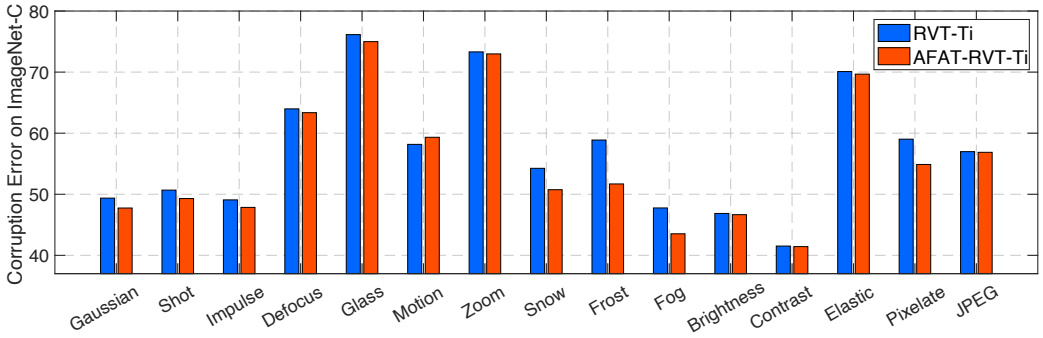

Figure 7: Comparisons of corruption error (the lower, the better) on individual corruption type of ImageNet-C between RVT-Ti and AFAT-RVT-Ti. Our AFAT model outperforms the RVT baseline model on most of the corruption types.

D More Visualization Results of Attention Stability

In the main paper, we have shown some examples to demonstrate the superiority of our AFATs in improving the stability of attention maps. Here, we additionally provide the visualization results of more examples. As highlighted by Figure 5 of the paper, our AFATs yield significantly more stable attentions, in terms of cosine similarity (Cos-Sim), than the baseline RVTs across the whole ImageNet dataset. Interestingly, we observe that the visual stability of attentions is highly correlated with Cos-Sim. To be specific, our AFAT often obtains very stable attentions on the examples with a Cos-Sim larger than 0.8, as shown in Figure 8. In addition, we also show some examples with the Cos-Sim lower than 0.8 in Figure 9.

As shown in Figure 8, following (Fu et al., 2022), we average the attention maps across all the attention heads in each layer and visualize the attention map for a query token, e.g., the center token highlighted by the red box. Given the randomly occluded examples, RVT often incurs significant changes in the attention maps. By contrast, our AFAT effectively preserves most of the regions with relatively high attention scores across layers. In addition, as occluding different patches may have different impacts, we show the distribution of Cos-Sim over 1000 randomly sampled occlusion masks for each image in Figure 8 (last column). Across different examples, our AFAT yields significantly better attention stability than RVT both qualitatively and quantitatively. In Figure 9, we also show two examples with a Cos-Sim lower than 0.8. These occluded examples distract the attentions of both RVT and our AFAT across layers. Nevertheless, our AFAT still yields better quantitative results than RVT. Overall, these results demonstrate the effectiveness of AFAT in improving the stability of self-attention.

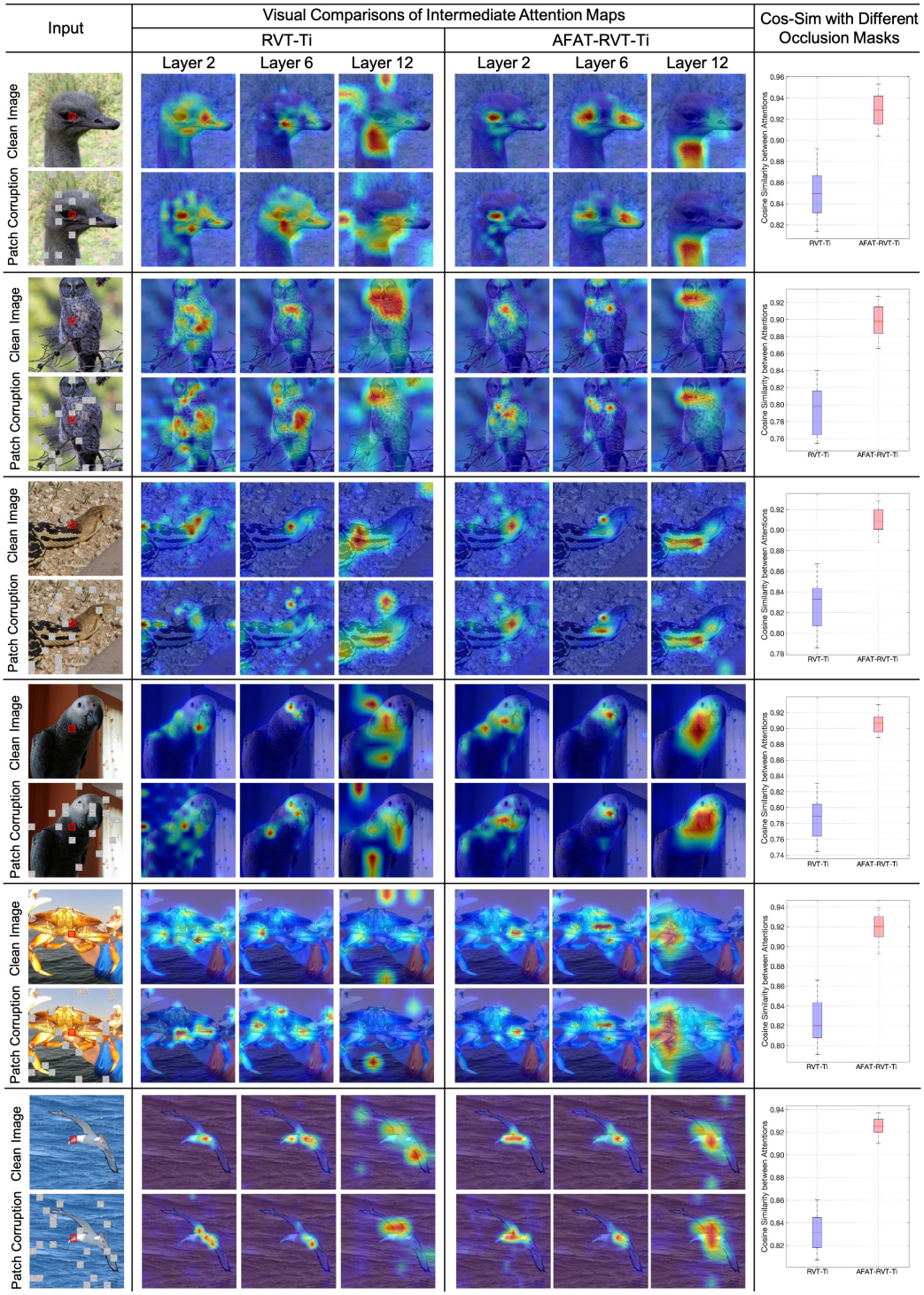

Figure 8: Comparisons of attention stability between RVT-Ti and our AFAT-RVT-Ti on the examples with relatively high cosine similarity (Cos-Sim). In the last column, we also investigate the impact of different occlusion masks (1000 random masks) on each example in terms of Cos-Sim. Clearly, our AFAT yields much more stable attention maps both qualitatively and quantitatively.

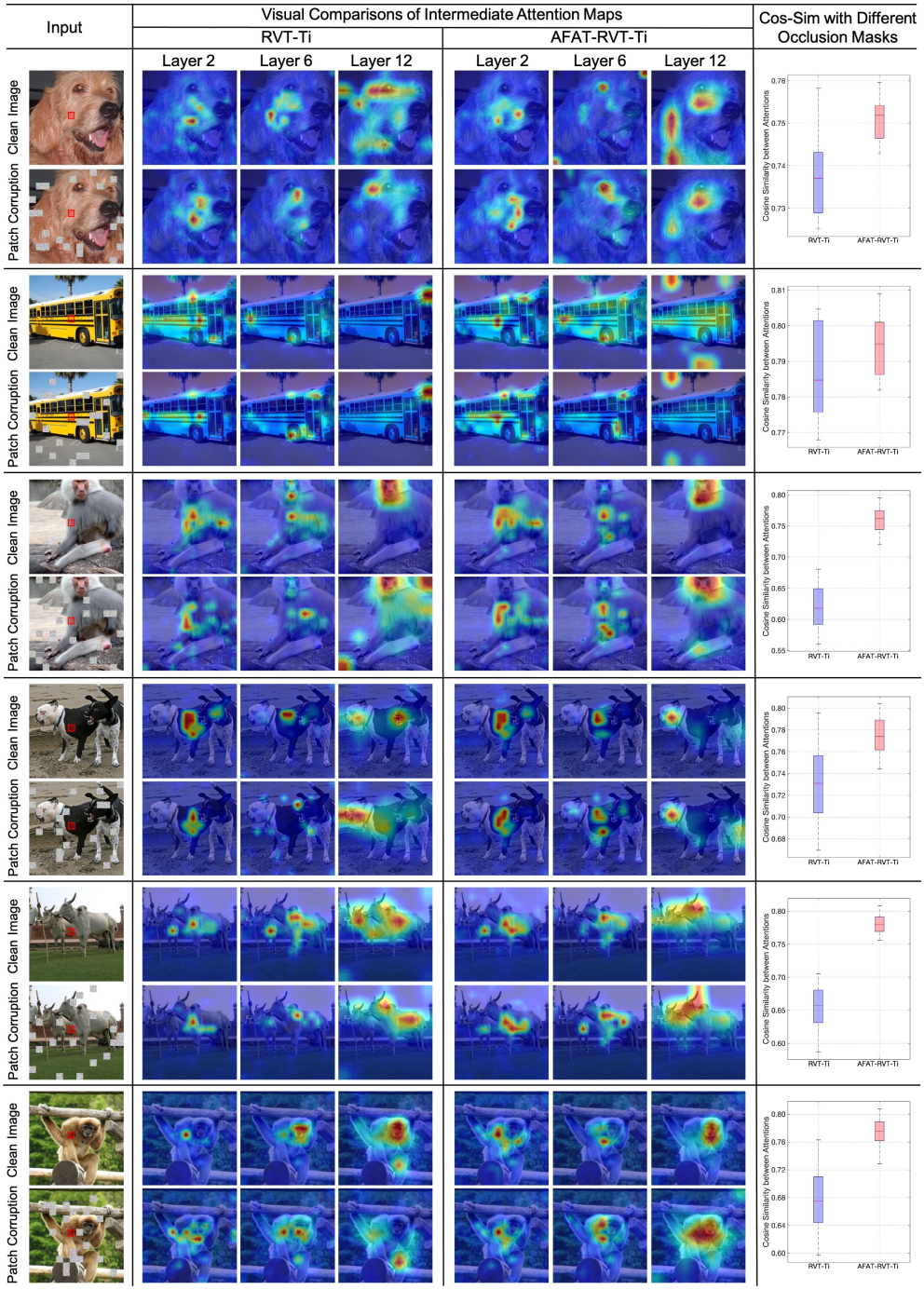

Figure 9: Comparisons of attention stability between RVT-Ti and our AFAT-RVT-Ti on the examples with relatively low cosine similarity (Cos-Sim). The occluded examples distract the attentions of both methods but our AFAT still yields higher Cos-Sim than RVT.